HOUSING DATA ANALYSIS
FOR HOMES IN KING COUNTY, WA

Dennis Trimarchi

2019-06-02

MOTIVATION & INITIAL ANALYSIS
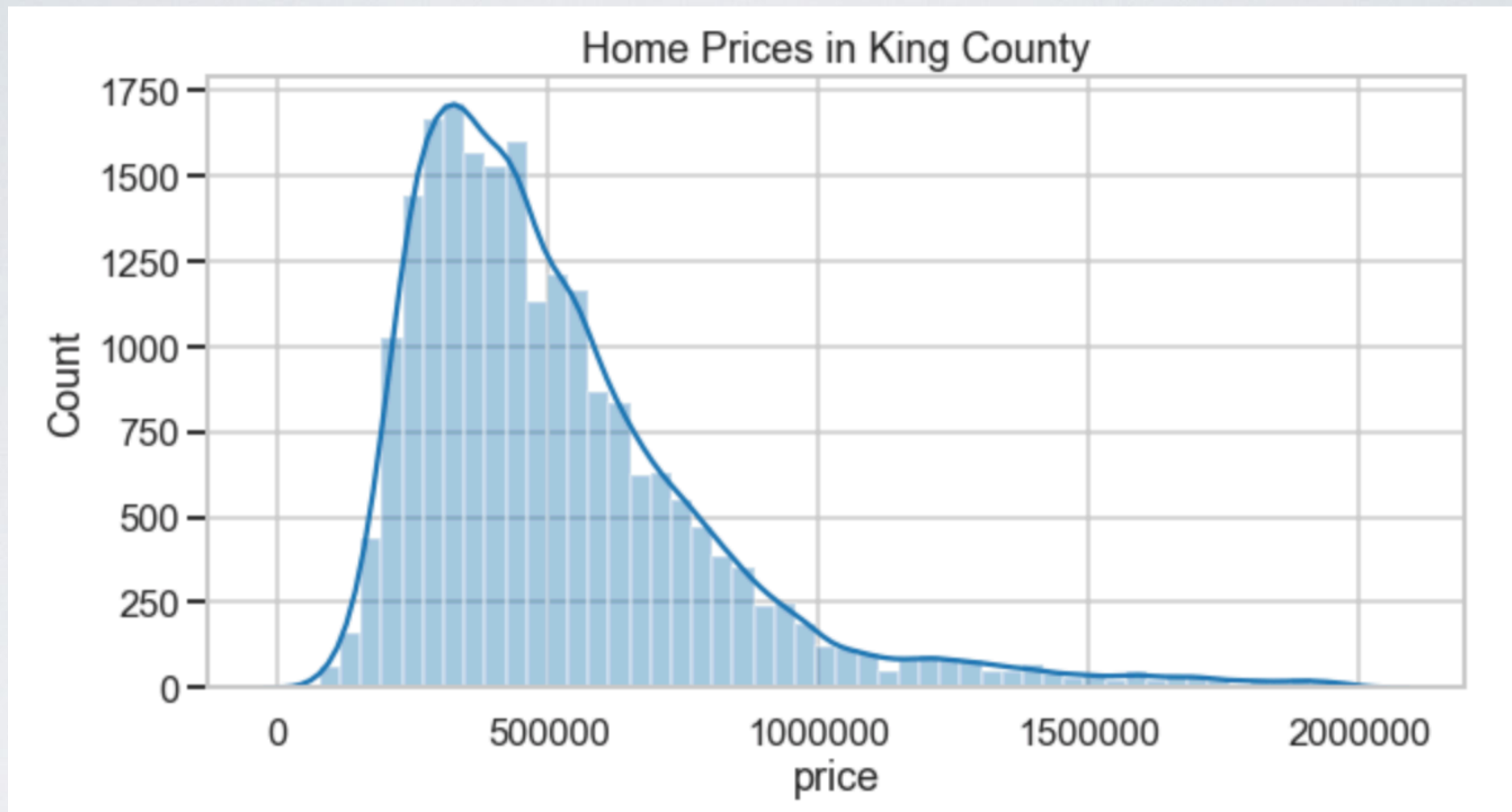
mean:
~ **$520K**

median:
~**$450K**

std dev:
~**$286K**

Home Prices in King County

- **Making predictions is useful**! Assessors make a living off of predicting home values. We want to see if an algorithm can be developed to accurately predict a home's price in King County, WA.

- A reliable set of data is needed: Removed 634 out of 21597 records that had missing / spurious data. **Removed homes valued over $2M** because they are not comparable to the majority of homes. Even at the $2M cutoff, the distribution is still skewed high.
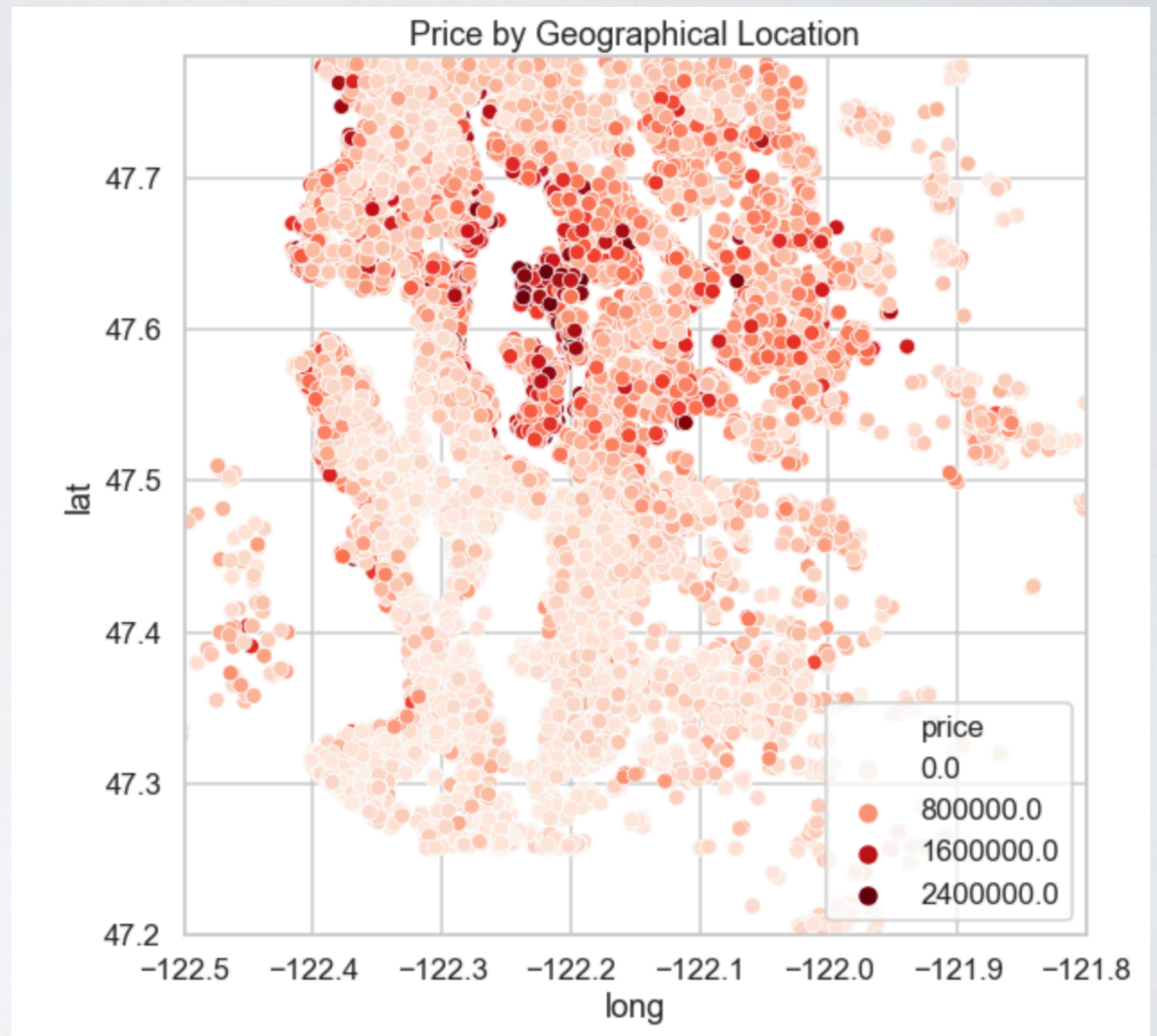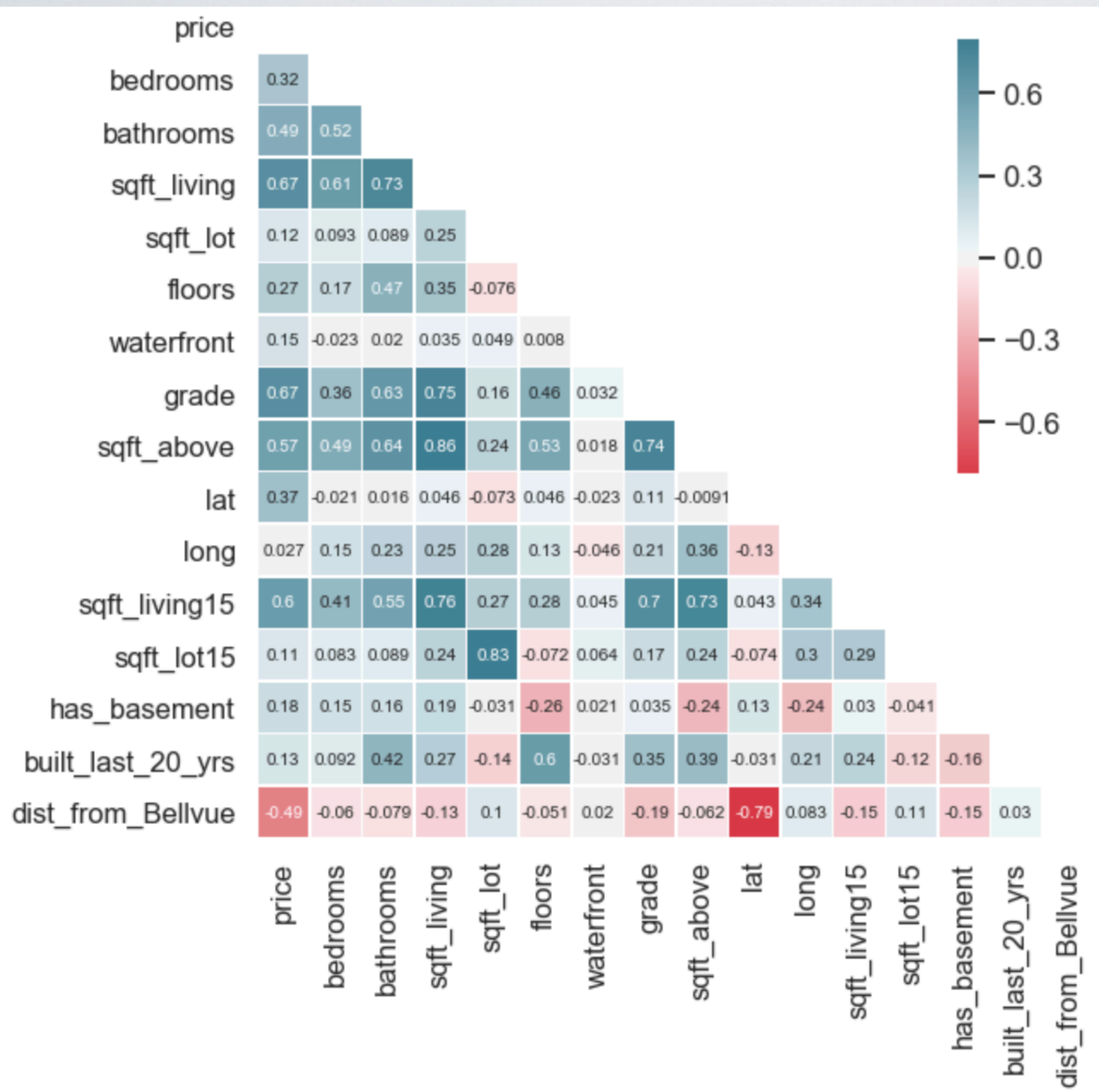
GEOGRAPHY

Some expensive homes in King County are <u>clustered</u> in a general area.

The darkest area of the map, and the center of a region of more expensive homes is <u>Bellevue, WA.</u>

Having a home in this region will impact price.

CORRELATION MATRIX

Price has strongest <u>positive</u> correlation with:
*sqft_living,
grade,
sqft_living15,
sqft_above,
bathrooms*

Price has strongest <u>negative</u> correlation with:
dist_from_Bellvue

Many of the features are correlated with each other in addition to price. This is bad model development as independent variables should be independent.
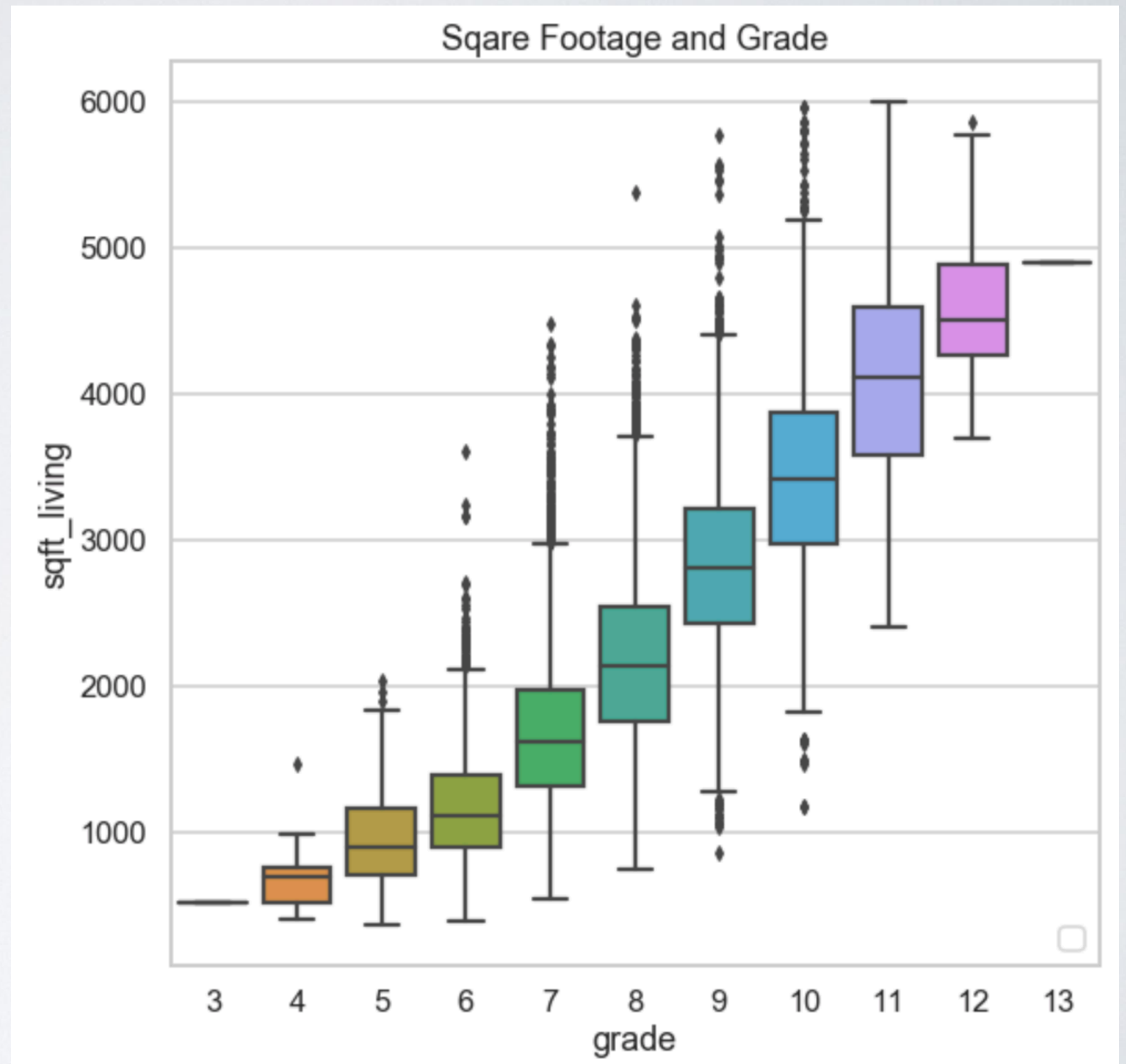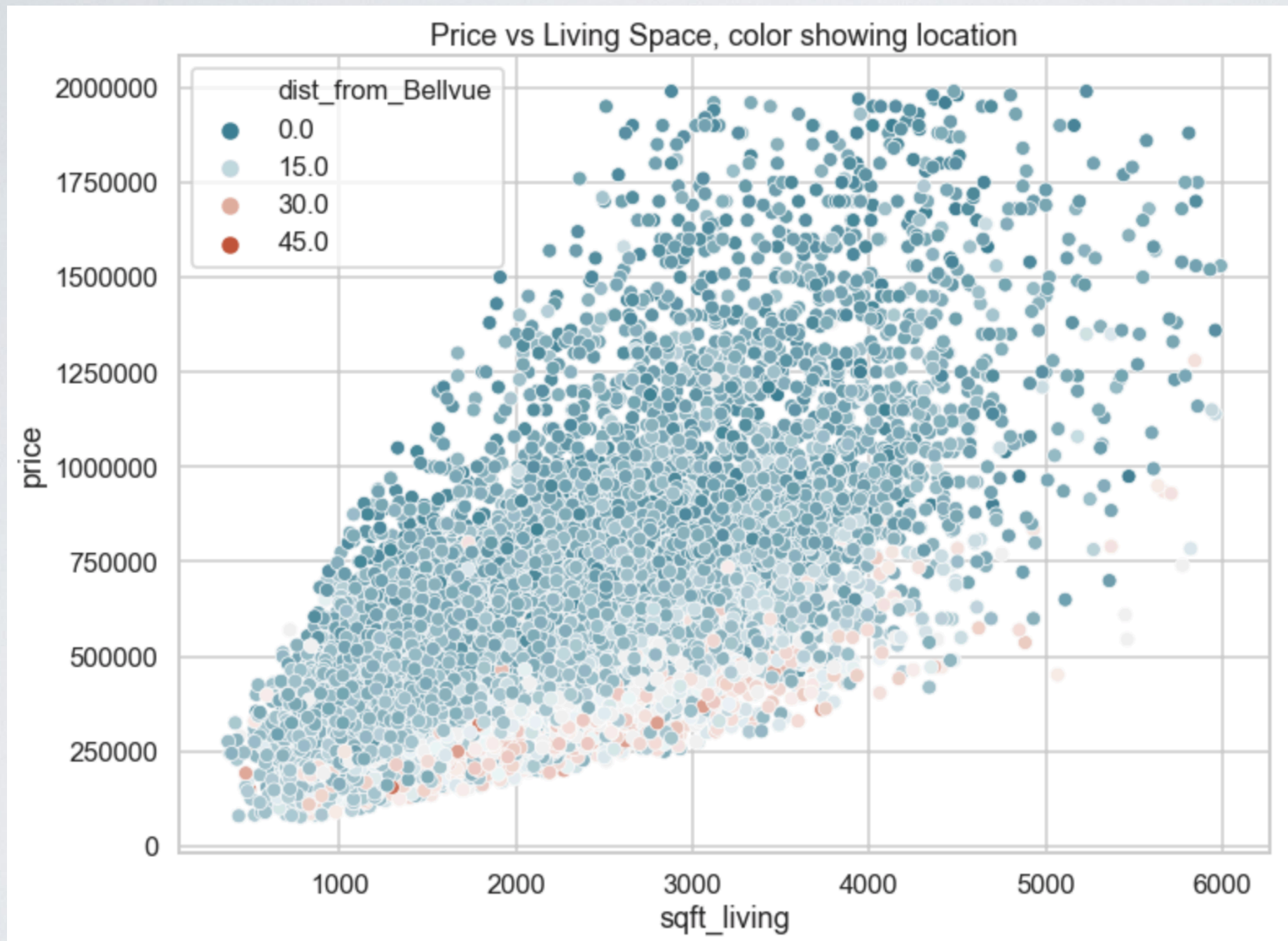
SQ FT LIVING & GRADE

These two fields are basically providing the same information. —Higher grades are given to larger homes. This is supported by the high correlation coefficient between the two features.

grade / average price
3     $262K
4     $212K
5     $247K
6     $301K
7     $401K
8     $540K
9     $762K
10    $993K
11    $1,209K
12    $1,541K
13    $1,780K

Sqare Footage and Grade

LIVING SPACE VS PRICE

Price vs Living Space, color showing location

Clear positive relationship between **price and living space**.

Also, distance from Bellevue, WA shown through the use of color.

As you can see, the further distances are clustered at the bottom of the distribution indicating lower price for being farther away.

ALGORITHM & DISCUSSION

Modeled Price as a function of sqft_living and dist_from_Bellvue. These particular features were selected based on high correlation to price and low correlation with each other. This makes for a stronger model.

$$price = 202.80 * sqft\_living - 196,800 * \ln(dist\_from\_Bellvue) + 550,300$$

The Model is simple in that it only contains two features and it uses information that should be readily available for any home in King County: square footage, and location.

There are some limitations in this model. Firstly, it can only be applied to data for King County. Homes across the state in other counties could not realistically use distance from Bellevue, WA as a way to predict home price.

Secondly, the algorithm only explains 61% of the variation in home price for the dataset. Other factors come into play when it comes to price.

FURTHER WORK

- Some additional information gathering would improve this model.

- In the Seattle area waterfront property is everywhere, and it would be nice to see if having a waterfront view impacts price. Unfortunately, for many of the homes this information was not provided, as such this feature was not included in the model.

- sqft_living was chosen over grade even though both features are highly correlated with price. With more data, it's quite possible that grade could be a better predictor. Also, it would be beneficial to know how the details of the King County grading system works.

- Another feature that could be investigated further is lot size. The data seemed to indicate a - city vs. rural split.

BACKUP SLIDES

WATERFRONT

Waterfront vs Price
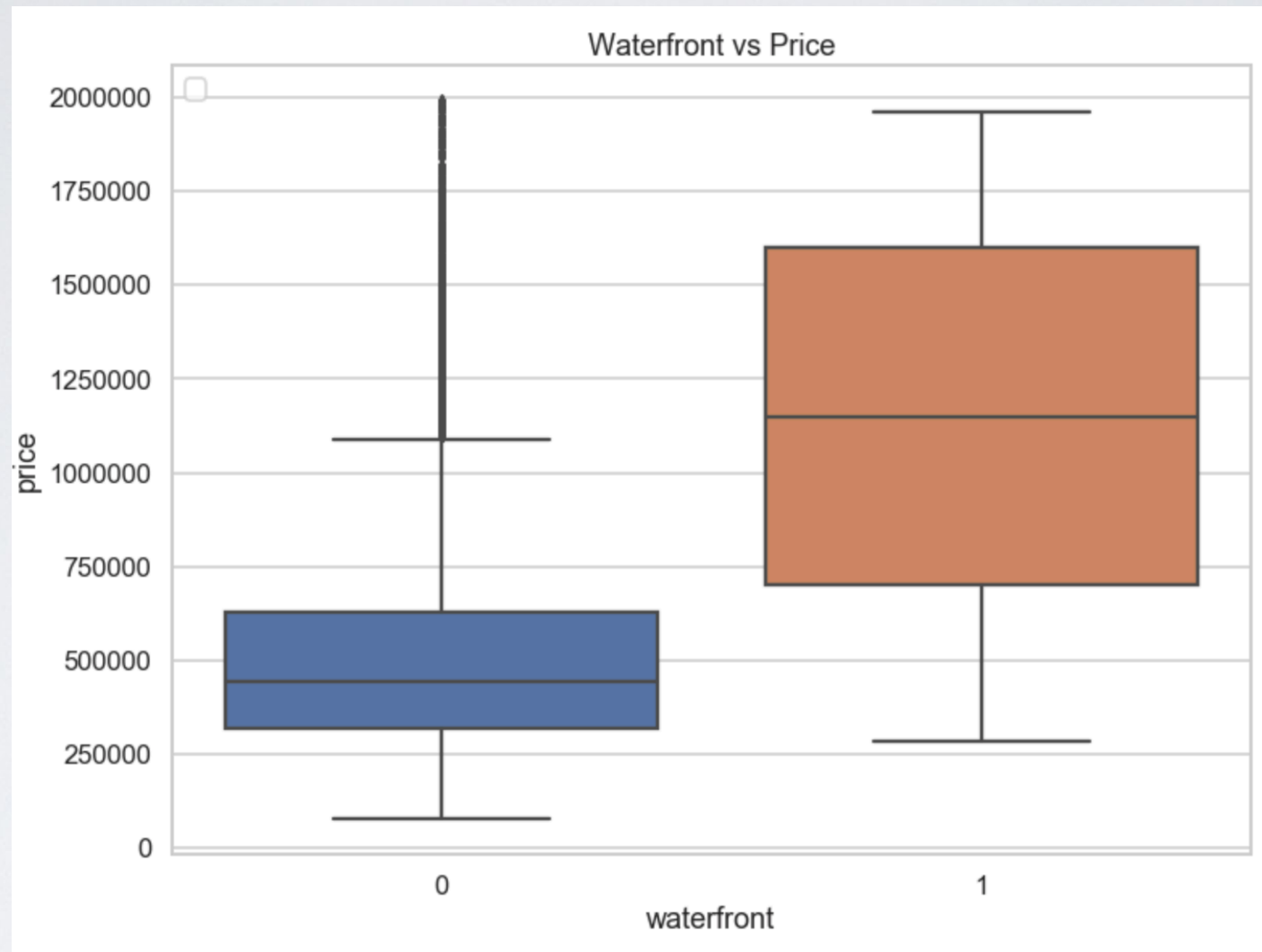
- Clear difference in price for waterfront properties.

LOT SIZE

- Having neighbors with large lot sizes (sqft_lot15) could indicate more rural areas.